Machine learning HW#4

Jiale Shi

1 Exponential family

A. The exponential family of distributions over y given parameter θ is given as

$$p(y|\theta) = h(y) \exp\{y\theta - A(\theta)\} \qquad \theta = \eta \qquad (1)$$

Present each of the following distributions in the exponential family form. Identify the relevant components necessary for use in a GLM: (i) the canonical parameter θ, (ii) $h(y)$, (iii) $A(\theta)$. Show your work

(a) Normal distribution

(b) Binomial distribution

(c) Poisson distribution

(d) Gamma distribution with distribution function

$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} e^{-y\nu/\mu}, \quad y > 0 \qquad (2)$$

(e) Inverse Gamma distribution with density function

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left\{-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right\}, \quad y, \mu\lambda > 0 \qquad (3)$$

(a) Normal distribution

$$P(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} = \exp\left\{-\frac{y^2}{2\sigma^2} + \frac{2\mu}{2\sigma^2} y - \underbrace{\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)}_{A(\theta)}\right\}$$

$$u(y) = (y, y^2)^T,$$

$$\theta = \left(\frac{2\mu}{2\sigma^2}, -\frac{1}{2\sigma^2}\right)^T, \quad h(y) = 1$$

$$A(\theta) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)$$

(b) Binomial distribution with known number of trails n.

$$f(y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y \in \{0, 1, 2, \cdots, n\}$$

$$f(y) = \binom{n}{y} \exp\left(y \log\left(\frac{p}{1-p}\right) + n \log(1-p) \right)$$

$$\theta = \log \frac{p}{1-p}, \quad u(y) = y$$

$$h(y) = \binom{n}{y}$$

$$A(\theta) = n \log(1-p)$$

(c) Poisson distribution

$$p(y|\lambda) = \text{Poisson}(y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = \underbrace{\frac{1}{y!}}_{h(y)} \exp(\underbrace{y}_{u(y)} \underbrace{\ln\lambda}_{\theta} - \underbrace{\lambda}_{A(\theta)})$$

$\theta = \ln\lambda$

$h(y) = \frac{1}{y!}$

$A(\theta) = -\lambda$

(d) Gaussian distribution with distribution function

$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} (y)^{\nu-1} e^{-y\frac{\nu}{\mu}}, \quad y>0$$

$$= \exp\left[(\nu-1)\ln y - y\frac{\nu}{\mu} + \ln\left(\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu}\right)\right]$$

$$= \exp\left[\begin{bmatrix}\ln y, & y\end{bmatrix}\begin{bmatrix}\nu-1 \\ -\frac{\nu}{\mu}\end{bmatrix} + \ln\left(\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu}\right)\right]$$

$$u(y) = [\ln y, y]^T$$

$$\theta = \begin{bmatrix}\nu-1 \\ -\frac{\nu}{\mu}\end{bmatrix}$$

$$h(y) = 1$$

$$A(\theta) = \ln\left[\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu}\right]$$

(e) Inverse Gamma distribution with density function

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left\{ - \frac{\lambda (y-\mu)^2}{2\mu^2 y} \right\}$$

$$= \sqrt{\frac{1}{y^3}} \exp\left\{ - \frac{\lambda (y-\mu)^2}{2\mu^2 y} + \frac{1}{2} \ln\left(\frac{\lambda}{2\pi}\right) \right\}$$

$$= \sqrt{\frac{1}{y^3}} \exp\left\{ - \frac{\lambda}{2\mu^2} \frac{y^2 - 2\mu y + \mu^2}{y} + \frac{1}{2} \ln\left(\frac{\lambda}{2\pi}\right) \right\}$$

$$= \sqrt{\frac{1}{y^3}} \exp\left[- \frac{\lambda}{2\mu^2} y + \frac{\lambda}{\mu} - \frac{\lambda}{2} \frac{1}{y} + \frac{1}{2} \ln\left(\frac{\lambda}{2\pi}\right) \right]$$

$$h(y) = \sqrt{\frac{1}{y^3}}$$

$$u(y) = \left[y, \frac{1}{y} \right]^T$$

$$\theta = \left[- \frac{\lambda}{2\mu^2}, - \frac{\lambda}{2} \right]$$

$$A(\theta) = \frac{\lambda}{\mu} + \frac{1}{2} \ln\left(\frac{\lambda}{2\pi}\right)$$

B. Data are generated for the exponential distribution with density $f(y) = \lambda \exp(\lambda)$ where $\lambda, y > 0$. The distribution is a member of the exponential family was shown before.

$f(y) = \lambda \exp(-\lambda y)$

(a) Identify the specific form of θ, $h(y)$ and $A(\theta)$ for the exponential distribution.

(b) Whats the canonical link for a generalized linear model (GLM) with a response following the exponential distribution?

(c) Identify a practical difficulty that may arise when using the canonical link in this instance

(a) $f(y) = \lambda \exp(-\lambda y)$

$= \exp(-y\lambda + \ln\lambda)$

$\theta = -\lambda$

$h(y) = 1$

$u(y) = y$

$A(\theta) = \ln\lambda$

(b) the canonical link is

$\mu = A'(\theta) = \frac{1}{\lambda}$

$\theta = -\lambda = -\frac{1}{\mu}$ Negative Inverse

(c) In the cases of exponential distribution, the domain of the canonical link function is not the same as the permitted range of the mean. In particular, the linear predictor may be positive, which would give an impossible negative mean. when maximizing the likelihood, precautions must be taken to avoid this. An alternative is to use a nonconanical function.

C. The Conway-Maxwell Poisson distribution has the probability function

$$P(Y=y)=\frac{\lambda^y}{(y!)^\nu}\frac{1}{Z(\lambda,\nu)},\quad y=0,1,2,\ldots \tag{4}$$

where

$$Z(\lambda,\nu)=\sum_{i=1}^{\infty}\frac{\lambda^i}{(i!)^\nu} \tag{5}$$

(a) Place this distribution in an exponential family form with respect to both parameters, and identify all the relevant components

(b) In statistics, overdispersion is the presence of greater variability (statistical dispersion) in a data set than would be expected based on a given statistical model. Explain why this distribution can be used to model overdispersion for count data.

(a) $\frac{\lambda^y}{(y!)^\nu}\frac{1}{Z(\lambda,\nu)}$

$= \exp\left(y\ln\lambda + \nu \ln y! - \ln Z(\lambda,\nu)\right)$

$h(y)=1$

$\theta = (\ln\lambda, \nu)^T$

$u(y) = (y, \ln y!)^T$

$A(\theta) = -\ln Z(\lambda,\nu)$

(b)

Mean: $E[Y] = \frac{\partial \ln Z}{\partial \ln\lambda} \approx \lambda^{\frac{1}{\nu}} + \frac{1}{2\nu} - \frac{1}{2}$

variance: $V[Y] = \frac{\partial^2 \ln Z}{\partial^2 \ln\lambda} = \frac{1}{\nu}\lambda^{1/\nu}$

$D[Y] = \frac{V[Y]}{E[Y]} \approx \frac{1}{\nu}$

when $\nu < 1$, $D(Y) > 1$

This is a result of overdispersed data.

when $\nu = 1$, $D(Y) = 1$. Poisson model.

when $\nu > 1$, $D(Y) < 1$

This is a result of underdispersed data.

2 Generalized linear model - Probit regression

In this homework, we consider the Bank data set problem discussed in detail in Chapter 4 (Generalized Linear Models) of the Bayesian Core book by Jean-Michel Marin, and Christian P. Robert. The data set can be downloaded from this link. In this data set, we have measurements on 100 genuine Swiss banknotes and 100 counterfeit ones. The response variable y is thus the status of the banknote, where 0 stands for genuine and 1 stands for counterfeit, while the explanatory factors are the length of the bill x_1, the width of the left edge x_2, the width of the right edge x3, and the bottom margin width x_4, all expressed in millimeters. We want a probabilistic model that predicts the type of banknote (i.e., that detects counterfeit banknotes) based on the four measurements above. In this context, do the following:

A. In GLMs, we have

$$\boldsymbol{y}|\mathbf{x}, \boldsymbol{w} \sim f\left(y|\mathbf{x}^T \boldsymbol{w}\right) \tag{6}$$

The above model is defined by two functions – a conditional density f of y given x that belongs to the exponential family and and that is parameterized by an expectation parameter $\mu = \mu(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$, and a a link function g that relates the mean $\mu = \mu(\mathbf{x})$ of f and the covariate vector, $\mathbf{x}$ as $g(\mu) = \mathbf{x}^T \boldsymbol{w}$ One of the popular link function is the probit link function $g(\mu_i) = \Phi^{-1}(\mu_i)$ where Φ is the standard normal cdf. The corresponding likelihood is given as

$$l(\boldsymbol{w}|\boldsymbol{y}, \mathbf{x}) \propto \prod_{i=1}^{n} \Phi\left(\mathbf{x}^{iT} \boldsymbol{w}\right)^{y_i} \left[1 - \Phi\left(\mathbf{x}^{iT} \boldsymbol{w}\right)\right]^{1-y_i} \tag{7}$$

Considering non-informative G-priors for the weights

$$\boldsymbol{w}|\sigma^2, \mathbf{x} \sim \mathcal{N}_k\left(0_k, \sigma^2 \left(\mathbf{x}^T \mathbf{x}\right)^{-1}\right), \tag{8}$$

and

$$\pi\left(\sigma^2|\boldsymbol{x}\right) \propto \sigma^{-3/2}, \tag{9}$$

compute an expression for the posterior.

$w|\alpha^2, X \sim \mathcal{N}_k\left(0_k, \alpha^2 (X^TX)^{-1}\right)$

$\pi(\alpha^2|X) \propto \alpha^{-\frac{3}{2}}$

Integrating out α^2 in this joint prior then leads to

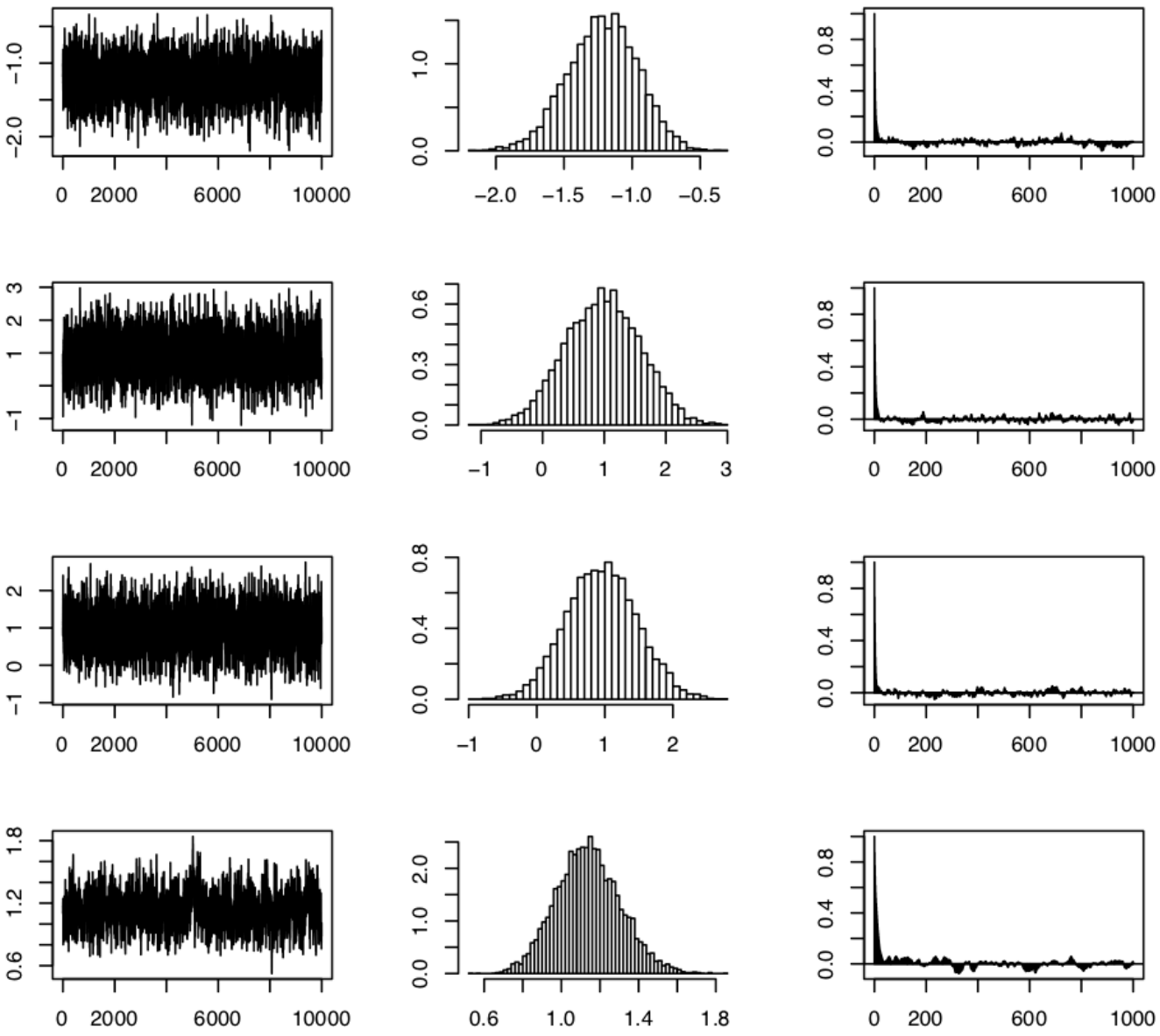
$\pi(w|X) \propto |X^TX|^{1/2} \Gamma((2k-1)/4) (w^T(X^TX)w)^{-(2k-1)/4} \pi^{-k/2}$

the corresponding posterior distribution of w is

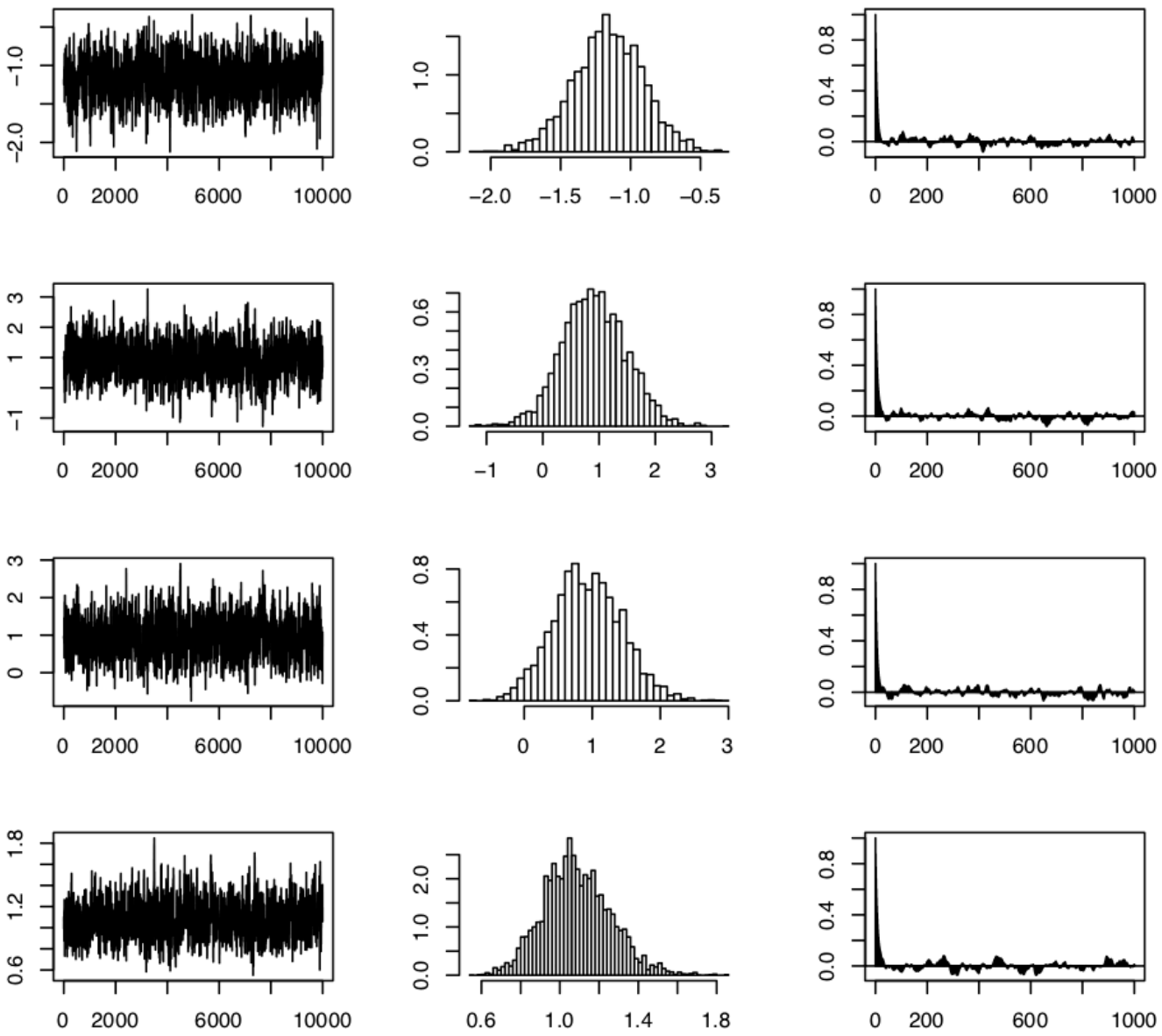
$\pi(w|y,X) = \pi(w|X) \cdot l(w|y,X)$

$\propto |X^TX|^{\frac{1}{2}} \Gamma((2k-1)/4) (w^T(X^TX)w)^{-(2k-1)/4} \pi^{-k/2} \times \prod_{i=1}^{n} \Phi(x^{iT}w)^{y_i} [1-\Phi(x^{iT}w)]^{1-y_i}$

B. In part A of this problem, you will see that it is not possible to provide analytical expression for the posterior. Therefore, one has to compute the posterior approximately. In this context, one option is to use the Metropolis Hasting (MH) algorithm as described in Algorithm 4.2 of [1]. Use MH algorithm to compute the posterior distribution of the weights (Similar to Fig. 4.5 in the book mentioned above). Run MH for 10,000 iteration. Consider the first 1000 iterations to be the burn-in period. Plot the histogram of the weights.



C. As an extension of the part B, write a code for drawing samples from the predictive distribution of y^*, $p(y^*|x^*)$. Plot histogram of the predictive distribution of y^* for a randomly generated sample of $\boldsymbol{x}^*$ (within the bound of the data set).



D. Computing marginal distribution of $\boldsymbol{y}$ is important to provide approximations to the Bayes factor. Unfortunately, this can not be computed in a closed form.

Provide an importance sampling based approximation to for the marginal distribution of $\boldsymbol{y}$, $p(\boldsymbol{y})$. Write a code for drawing samples from the marginal distribution of $\boldsymbol{y}$. Use it to compute the Bayes factor corresponding to the null hypothesis, $H_0 : w_i = 0$.

bf0probit	126.331869802136
bf1probit	3582.79455009591
bf2probit	0.0993327996344047
bf3probit	0.13719715182382
bf4probit	237509022530660
mkprob	num [1:4] -1.162 0.914 0.928 1.083

```
> log10(bf0probit)
[1] 2.101513
> mkprob
[1] -1.1615231  0.9143048  0.9283489  1.0825798
> diag(vkprob)
[1] 0.06469255 0.32737918 0.25016770 0.02901700
> log10(c(bf1probit,bf2probit,bf3probit,bf4probit))
[1]  3.5542219 -1.0029073 -0.8626549 14.3756801
>
```

3 Generalized linear model - Logit regression

Consider the same data set as previous problem and perform the following tasks

A. An alternative to the probit regression model is the logit regression model. In this case, the link function $g(\mu_i) = \log(\mu_i/(1-\mu_i))$. For this case, compute the posterior. For this model, the likelihood can be written as

$$l(\boldsymbol{w}|\boldsymbol{y}, \mathbf{x}) = \frac{\exp\left\{\sum_{i=1}^{n} y_i \mathbf{x}^{iT}\boldsymbol{w}\right\}}{\prod_{i=1}^{n}\left[1+\exp\left(\mathbf{x}^{iT}\boldsymbol{w}\right)\right]} \tag{10}$$

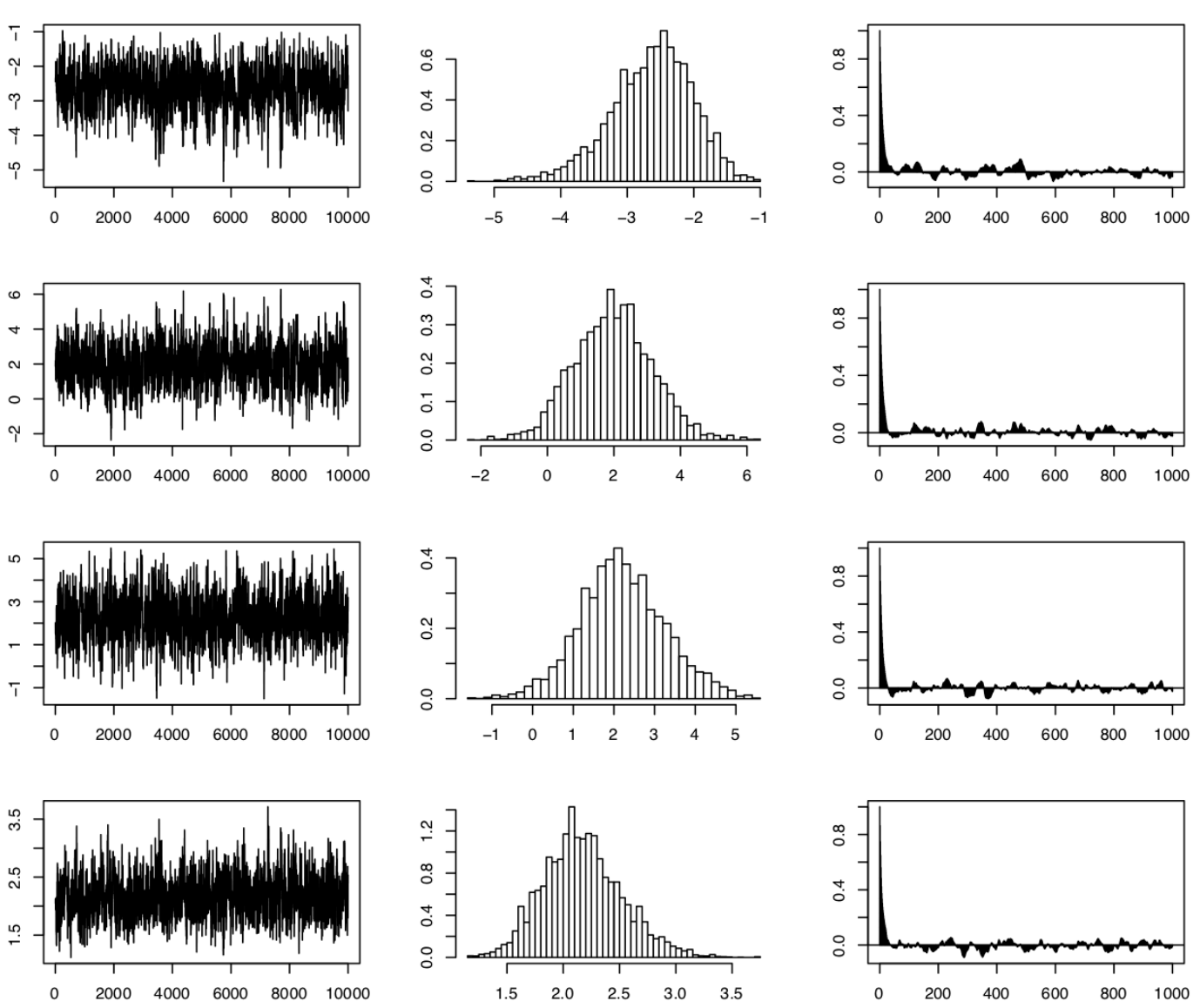
The prior is similar to that defined in the previous problem. Compute an expression for the posterior of $\boldsymbol{w}$.

the corresponding posterior distribution of w is

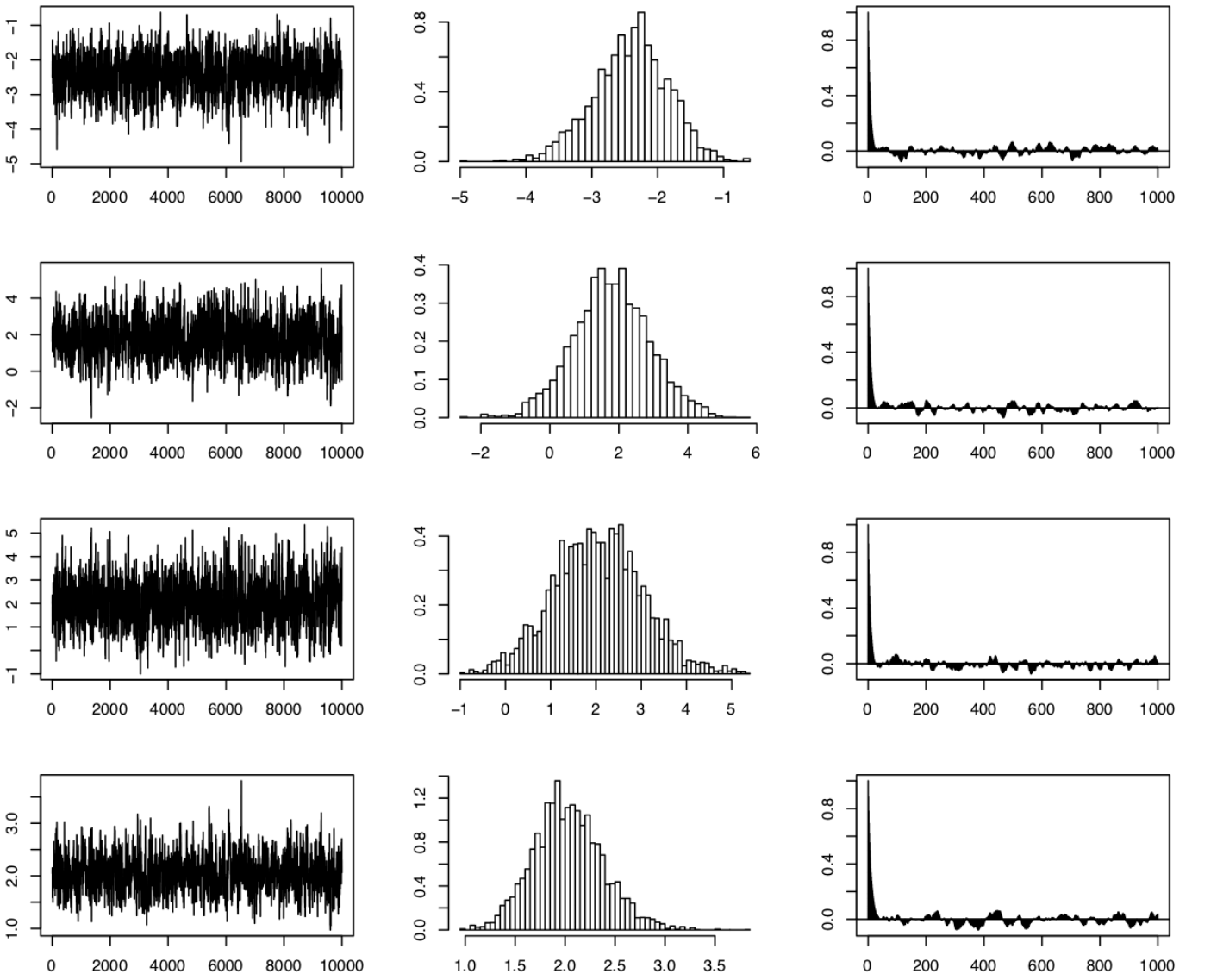
$$\pi(w|y,X) = \pi(w|X)\cdot l(w|y,X)$$

$$\propto |X^TX|^{\frac{1}{2}}\Gamma((2k-1)/4)(w^T(X^TX)w)^{-(2k-1)/4}\pi^{-k/2}\frac{\exp\left\{\sum_{i=1}^{n} y_i x^{iT}w\right\}}{\prod_{i=1}^{n}[1+\exp(x^{iT}w)]}$$

B. n part A of this problem, you will see that it is not possible to provide analytical expression for the posterior. Therefore, one has to compute the posterior approximately. In this context, one option is to use the Metropolis Hasting (MH) algorithm as described in Algorithm 4.2 of [1]. Use MH algorithm to compute the posterior distribution of the weights (Similar to Fig. 4.5 in the book mentioned above). Run MH for 10,000 iteration. Consider the first 1000 iterations to be the burn-in period. Plot the histogram of the weights.



C. As an extension of the part B, write a code for drawing samples from the predictive distribution of y^*, $p\left(y^*|x^*\right)$. Plot histogram of the predictive distribution of y^* for a randomly generated sample of $\boldsymbol{x}^*$ (within the bound of the data set).



4 K-means algorithm

For the yeast gene expression data provided at this link, use K-means algorithm to compute the cluster centers. Assume, you have 16 clusters

Use K-mean algorithm.

16 clusters. $K = 16$

apply the kmeansFit.m

the mu is the cluster center

```
mu =

  Columns 1 through 8

    0.0016   -0.0727    0.0459   -0.7518   -0.0714    0.0222   -0.2883   -0.2651
    0.2023    0.3552    0.4445   -1.3334    0.1263    0.0866   -0.1262   -0.2101
    0.3392    0.9663    1.1542    0.0642    0.4595    0.2441    0.0251   -0.0801
    1.0835    1.9276    1.4103    0.2308    0.9447    0.4602   -0.0344    0.2923
    1.3619    2.1197    1.7363   -0.0129    1.4879    0.9894   -0.1936    0.4385
    3.4361    3.3982    1.5540    2.2176    2.9045    1.3402    1.3628    2.2083
    3.2026    3.4223    1.7206    2.3027    2.0180    2.3962    2.6782    1.4003

  Columns 9 through 16

   -1.0794    0.2478    0.1730    0.1152    0.2637   -0.0357    0.0504   -0.0225
   -1.7873    0.1731    0.2980   -0.0442    0.2106    0.0788   -0.0950    0.1692
   -0.0114    0.0731   -0.1570   -0.3280    0.3548    0.3370   -0.3775    0.0921
   -0.2933   -0.6549   -0.4421   -0.9054   -0.1894    0.7120   -0.7819    0.6612
   -1.5741   -0.7432   -0.1376   -1.0533   -0.3825    0.8668   -0.8369    1.0323
    1.1025   -1.8920   -1.7895   -2.4392   -1.5032    2.1677   -2.2549    2.2415
    0.8284   -1.9879   -2.2460   -2.4084   -2.0546    1.9705   -0.5464    2.7439
```

5 Gaussian mixture, expectation maximization and mixture of experts

A. Show the student's t distribution can be represented as infinite mixture of Gaussian. For simplicity, assume one dimensional-distribution.

Gaussian distribution

$$g_{\mu,\tau}(x) = \frac{\sqrt{\tau}}{\sqrt{2\pi}} e^{-\frac{\tau(x-\mu)^2}{2}}, \quad \tau = \frac{1}{\sigma^2}$$

Gamma distribution

$$h_{\alpha,\beta}(\tau) = \frac{1}{\Gamma(\alpha)} e^{-\frac{\tau}{\beta}} \tau^{-1+\alpha} \beta^{-\alpha}$$

$$f_{\mu,\alpha,\beta}(x,\tau) = g_{\mu,\tau}(x)\, h_{\alpha,\beta}(\tau)$$

$$f_{\mu,\alpha,\beta}(x,\tau) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)\sqrt{2\pi}} e^{-\tau\left(\frac{(x-\mu)^2}{2}+\frac{1}{\beta}\right)} \tau^{-1/2+\alpha}$$

$$\int_0^{\infty} f_{\mu,\alpha,\beta}(x,\tau)\, d\tau$$

$$= \int_0^{\infty} f_{\mu,\alpha,\beta}(x,\tau) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)\sqrt{2\pi}} e^{-\tau\left(\frac{(x-\mu)^2}{2}+\frac{1}{\beta}\right)} \tau^{-1/2+\alpha} d\tau$$

$$= \frac{\sqrt{\beta}\,\Gamma\left(\alpha+\frac{1}{2}\right)}{\sqrt{2\pi}\,\Gamma(\alpha)} \frac{1}{\left(\frac{\beta}{2}(x-\mu)^2+1\right)^{\alpha+\frac{1}{2}}}$$

$$\alpha = \frac{\nu}{2}, \quad \frac{\beta}{2} = \frac{1}{\nu}$$

$$= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1+\frac{(x-\mu)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

B. It is possible to interpret the probit regression model as a latent variable model. In this setup, we associate each otem x_i with two utilities u_{0i} and u_{1i}, corresponding to the possible choices of $y_i = 0$ and $y_i = 1$. We then assume that the observed choice is whichever action has larger utility. Further details on this representation can be found in Section 9.4.2 of [2].

In this context, show how the probit regression, represented as latent variable model, can be trained using expectation maximization.

$P(y_i=1 \mid z_i) = \mathbb{I}(z_i > 0)$, where $z_i \sim N(w^T x_i, 1)$ is latent.

Although it is possible to fit probit regression model using gradient based methods, the EM-based approach has the advantage that it generalized to many other kinds of models.

The complete data log likelihood has the following form, assuming a $N(0, V_0)$ prior on w

$$\ell(z, w \mid V_0) = \log P(y \mid z) + \log N(z \mid Xw, I) + \log N(w \mid 0, V_0)$$

$$= \sum_i \log P(y_i \mid z_i) - \frac{1}{2}(z - Xw)^T(z - Xw) - \frac{1}{2} w^T V_0^{-1} w + \text{const}$$

The posterior in the E step is a truncated Gaussian

$$P(z_i \mid y_i, x_i, w) = \begin{cases} N(z_i \mid w^T x_i, 1)\mathbb{I}(z_i > 0), & \text{if } y_i = 1 \\ N(z_i \mid w^T x_i, 1)\mathbb{I}(z_i < 0), & \text{if } y_i = 0 \end{cases}$$

w only depends linearly on z, so we just need to compute $E[z_i \mid y_i, x_i, w]$

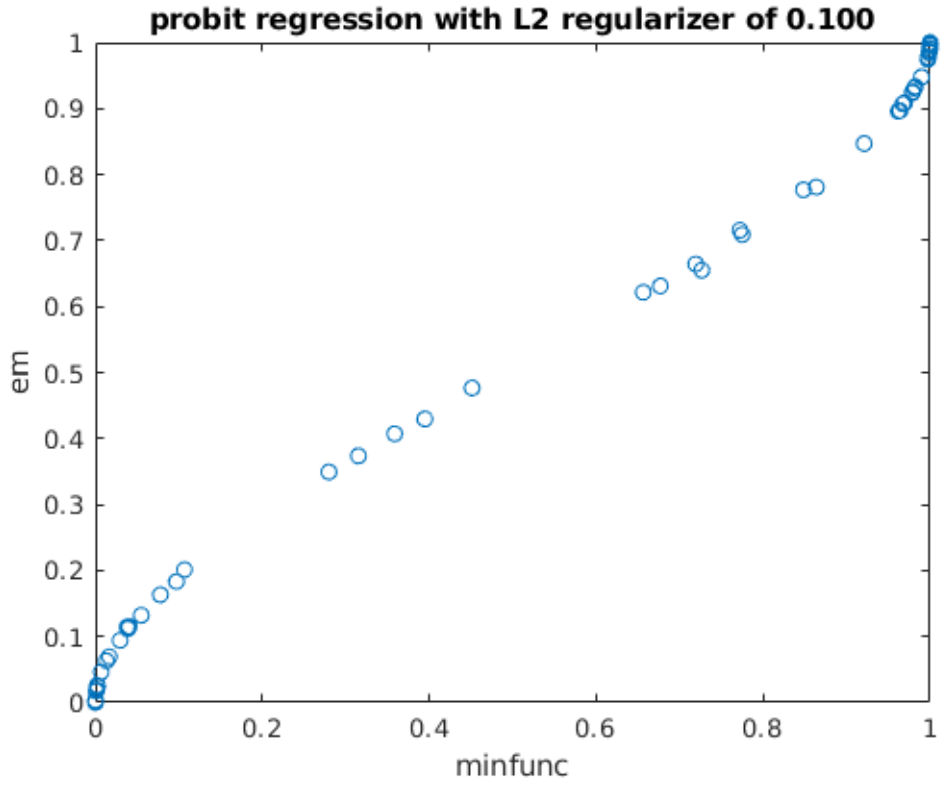
$$E[z_i \mid w, x_i] = \begin{cases} \mu_i + \dfrac{\phi(\mu_i)}{1 - \Phi(-\mu_i)} = \mu_i + \dfrac{\phi(\mu_i)}{\Phi(\mu_i)}, & \text{if } y_i = 1 \\ \mu_i - \dfrac{\phi(\mu_i)}{\Phi(-\mu_i)} = \mu_i + \dfrac{\phi(\mu_i)}{1 - \Phi(\mu_i)}, & \text{if } y_i = 0 \end{cases}$$

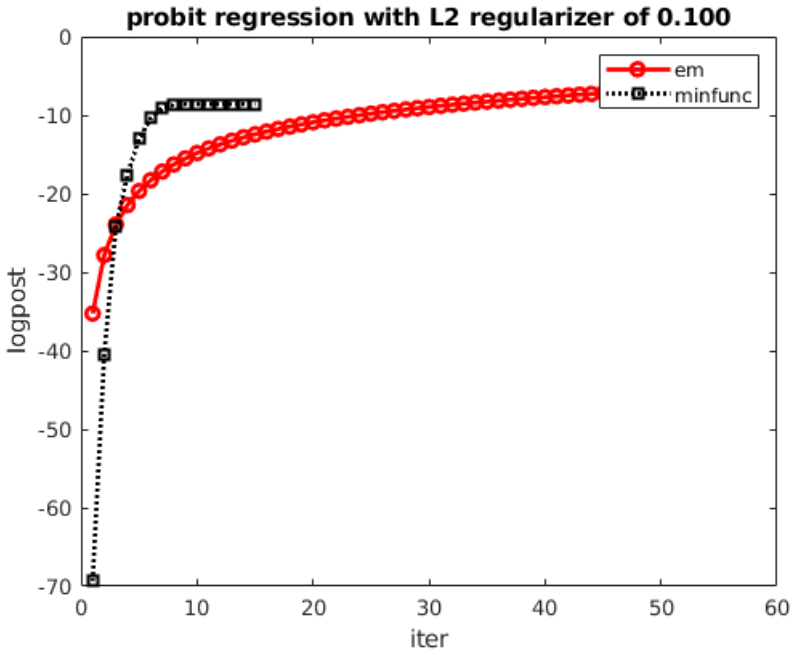
$$\mu_i = w^T x_i$$

In the M step, we estimate w using ridge regression, where $\mu = E[z]$ is the output we are trying to predict. Specifically, we have

$$\hat{w} = (V_0^{-1} + X^T X)^{-1} X^T \mu$$

C. For the data set given at this link, train a probit regression model by using expectation maximization. Compare with conventional probit regression model. Provide your observations.





The EM algorithm is simple, but can be much slower than direct gradient methods, from the figure above. This is because the posterior entropy in the E step is quite high, since we only observe that z is positive or negative, but are given no information from the likelihood about its magnitude.

D. Consider the linear regression model:

$$y = w_0 + w_1 x \tag{11}$$

For the given data set in this link, compute the unknown coefficients using the student's T model. This is the same problem as 1A of HW3. However, in this case, assume all the parameters to be unknown.

In this case, assume all the parameters to be unknown.
dof and sigma2 are unknown.

I run the linregRobustStudentFit(Xtrain, y)

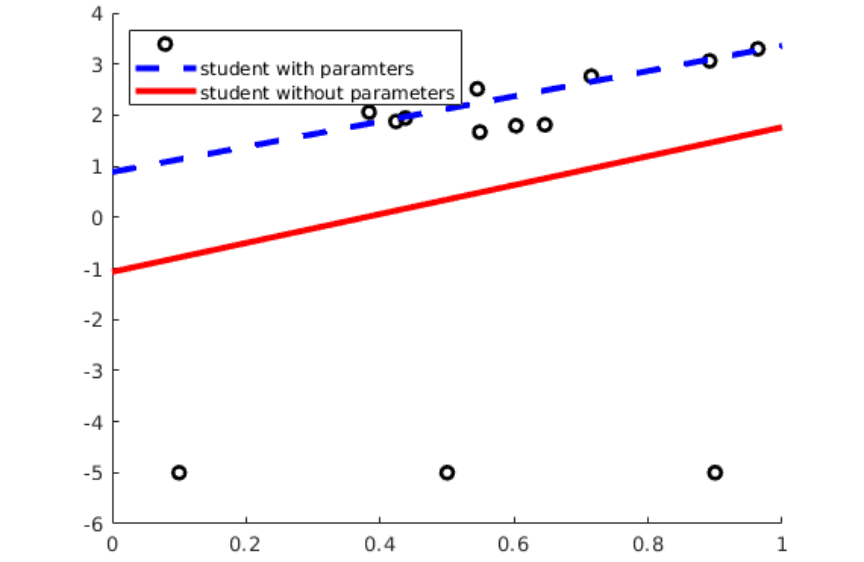
the results give the same dof and sigma2,
but different w0 and w1.

With parameters
dof and sigma2

w0	0.8858
w1	2.4730
dof	0.6296
sigma2	0.0688

without parameters
no dof and sigma2

w0	-1.0691
w1	2.8321
dof	0.6296
sigma2	0.0688



E. Often a single regression model may not be sufficient for tracking variability of a function in the overall problem domain. Under such circumstances, a good option is to use multiple regression model, each applied to a separate input space. We can model this by allowing the mixing weights and the mixture densities to be input-dependent. Such models are known as mixture of experts (MOE). Mixtures of experts are useful in solving inverse problems. These are problems where we have to invert a many-to-one mapping.

For the data-set provided this link, fit a mixture of experts with three experts. Plot the predictive mean and mode.

Consider each of the three experts to be a linear regression model of the form

$$y = w_0 + wx \tag{12}$$